COMPUTATIONS PERTAINING TO SOME SEQUENCE RECONSTRUCTION PROBLEMS IN IMMUNOGENOMICS

V. ARVIND RAMESHWAR

In this note, we consider classes of problems in sequence reconstruction that are motivated by immunogenomics. The branch of *personalized immunogenomics* seeks to derive the *germline genes* (or genes in the immunoglobin locus) of an individual from (several) samples of randomly selected immunoglobin genes (also called *antibody repertoire*). The sequences in the antibody repertoire, which can be seen as "traces," in the language of the computer science literature, have undergone "errors" in the form of *somatic hypermutations* (SHMs) or *clonal selection*, in order to modify the affinity of the antibodies to specific antigens. Hence, the task of obtaining the individual germline genes can be seen as an instance of a sequence reconstruction problem, given erroneous traces [1]. We refer the reader to the rich literature on classical sequence reconstruction problems, beginning with the works of Levenshtein [2, 3], and progressing to the more recent results on "trace reconstruction" over the deletion channel [4, 5, 6, 7, 8], and the references therein. Notably, this write-up considers channel models, inspired by problems in immunogenomics, which are quite different from the channels introducing "memoryless" errors, such as in the previous references.

Our chief interest will be in deriving asymptotic bounds on the number of traces (also called "trace complexity") required for reconstructing a (fixed or random) germline gene, with at most a fixed probability of error. Our analysis will be entirely based on channel models for this sequence reconstruction problem, which were introduced in [1, Sec. III]. While the work [1] set down the channel laws for the models introduced, it left open the question of analyzing bounds on the trace complexity for reliable reconstruction. Indeed, as we shall see, manipulations/reformulations of the channel laws directly give rise to simple algorithms, which in turn provide upper bounds on trace complexity, and lower bounds, via a standard information-theoretic argument. The main contributions of this write-up are hence the somewhat tedious computations required for arriving at these bounds on the trace complexity.

1. Channel Models and Preliminaries

In this section, we recapitulate the channel models presented in [1, Sec. III]. We focus exclusively on channel models for the task of reconstructing the D gene in the immunoglobin locus; more complicated models for reconstructing the V, D, and J genes in the immunoglobin locus, based on concatenations of outputs of channel models for the D gene can also be found in [1, Sec. III].

Let n be the length of the D gene, which is treated as a sequence $\mathbf{x} \in \{0, 1, 2, \ldots, q-1\}^n$, for some fixed size $q > 1$ of the alphabet $\mathcal{X} := \{0, 1, 2, \ldots, q-1\}$. We are interested in the following channel models:

(1) TrimSuffixAndExtend (or channel W_1): An integer $R \sim \text{Unif}([0 : n])$ is sampled and the last R symbols of $\mathbf{x}$ are replaced with symbols that are drawn i.i.d. uniformly from $\mathcal{X}$.
(2) TrimAndExtend (or channel W_2): A pair of non-negative integers (R_1, R_2) is drawn uniformly from all such pairs whose sum is at most n. The first R_1 symbols and the last R_2 symbols of $\mathbf{x}$ are replaced with symbols that are drawn i.i.d. uniformly from $\mathcal{X}$.
(3) SuffixExtend$_t$(TrimSuffix) (or channel W_3): An integer $R \sim \text{Unif}([0 : n])$ is sampled and the last R symbols of $\mathbf{x}$ are trimmed, giving rise to the trimmed sequence $\mathbf{x}'$. Next, for a fixed integer $t \geq 1$, an integer $E \sim \text{Unif}([0 : t])$ is sampled and E symbols that are drawn i.i.d. uniformly from $\mathcal{X}$ are appended at the end of $\mathbf{x}'$.

We believe that similar computations as those made for channels W_1–W_3 can also be carried out for channel models that allow for "mutations" or symbol flips, following corruptions by the channels above. As an additional piece of notation, given a channel W, we define

$$d_n(W) := \min_{\mathbf{u} \neq \mathbf{u}' \in \mathcal{X}^n} d_{\text{TV}} \left(W(\cdot|\mathbf{u}), W(\cdot|\mathbf{u}') \right)$$

to be the smallest total variational distance between channel transition probabilities corresponding to distinct channel input sequences. We now define a sequence reconstruction algorithm over a channel W.

Definition 1 (Sequence reconstruction algorithm)**.** *A sequence reconstruction algorithm A takes as input traces $\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathcal{X}^\star$, for some $N \geq 1$, each obtained independently via a channel W, and returns as output a sequence (gene) $\widehat{\mathbf{x}} \in \mathcal{X}^n$.*

It is clear that a sequence reconstruction algorithm is useful only if it returns the correct input sequence, with high probability. We now recapitulate the definitions of the average-case error probabilities of a sequence reconstruction algorithm and the average-case trace complexity. Recall the definition of the maximum a-posteriori probability (MAP) decoder MAP.

Definition 2 (Average-case error probability and trace complexity)**.** *The average-case error probability of a sequence reconstruction algorithm A over the channel W, given N traces, is*

$$P_{e,A}^{(N)} := \frac{1}{q^n} \cdot \sum_{\mathbf{x} \in \mathcal{X}^n} P_{e,A}^{(N)}(\mathbf{x}),$$

where

$$P_{e,A}^{(N)}(\mathbf{x}) := \sum_{(\mathbf{y}_1, \ldots, \mathbf{y}_N) \in \mathcal{X}^\star} \prod_{i=1}^{N} W(\mathbf{y}_i|\mathbf{x}) \cdot \mathbb{1}\{A(\mathbf{y}_1, \ldots, \mathbf{y}_N) \neq \mathbf{x}\}.$$

The average-case trace complexity, for a fixed error probability $\delta \in (0,1)$, is the smallest number of traces required on average by the optimal MAP decoder:

$$T_\delta(n) = \min\{N : P_{e,\text{MAP}}^{(N)} \leq \delta\}.$$

Definition 3 (Trace complexity given an input sequence)**.** *The trace complexity given an input sequence $\mathbf{x} \in \mathcal{X}^n$, for a fixed error probability $\delta \in (0,1)$, is given by*

$$T_\delta(n; \mathbf{x}) := \min\{N : P_{e,\text{MAP}}^{(N)}(\mathbf{x}) \leq \delta\}.$$

The following simple lemma is then immediate from the definitions above.

Lemma 1. *For the channels W_1 through W_4, we have that $T_\delta(n) = T_\delta(n; \mathbf{x})$, for all $\delta \in (0,1)$, $n \geq 1$, and $\mathbf{x} \in \mathcal{X}^n$.*

Proof. The proof directly follows from the fact that $P_{e,\text{MAP}}^{(N)}(\mathbf{x})$, for each of these channels and for any $\mathbf{x} \in \mathcal{X}^n$, is independent of $\mathbf{x}$, which in turn holds since the channel laws are dependent only on the errors introduced, and not on the input sequence. □

In what follows, we hence restrict attention to the setting where the all-zeros sequence is transmitted over channels W_1 through W_4, and obtain bounds on the quantity $T_\delta(n)$, via bounds on $T_\delta(n; \mathbf{0})$. The following lemma, which holds via standard arguments analogous to the discussion after [5, Thm. 1.2], will be useful to us.

Lemma 2. *For any channel W and for all $\delta \in (0,1)$, we have that the trace complexity $T_\delta(n) = \Omega\left(\frac{1}{d_n(W)}\right)$.*

We next introduce a simple algorithm, FINDMODE (see Algorithm 1), that returns the trace occurring most often. As we shall see, all our reconstruction algorithms are either FINDMODE itself, for a certain values of N, or are simple variants thereof.

Algorithm 1 Sequence reconstruction algorithm

1: **procedure** FINDMODE($\mathbf{y}_1, \ldots, \mathbf{y}_N$)
2: Return the sequence occurring most often in $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$.

2. TRIMSUFFIXANDEXTEND

In this section, we obtain bounds on the trace complexity, for error probability $\delta \in (0,1)$, over channel W_1. Given sequences $\mathbf{u}, \mathbf{y} \in \mathcal{X}^n$, we define $\ell(\mathbf{u}, \mathbf{y})$ to be the length of the longest common prefix of $\mathbf{u}$ and $\mathbf{y}$. First, we present a simple lemma, which is a restatement of the channel law of W_1 in [1, Sec. III-A].

Lemma 3. *For any $\mathbf{u}, \mathbf{y} \in \mathcal{X}^n$, we have*

$$W_1(\mathbf{y}|\mathbf{u}) = \frac{1}{(n+1)q^n} \cdot \left(\frac{q^{\ell(\mathbf{u},\mathbf{y})+1} - 1}{q-1} \right).$$

The characterization above directly gives rise to a simple sequence reconstruction algorithm, via FINDMODE.

Theorem 4. *For any $\delta \in (0,1)$, we have $T_\delta(n) \leq 2(n+1)^2 \cdot \left(\ln\left(\frac{1}{\delta}\right) + n \ln q\right)$.*

Proof. Recall from Lemma 1 that it suffices to consider the trace complexity $T_\delta(n; \mathbf{0})$. Now, via Lemma 3, we have that

$$W_1(\mathbf{y}|\mathbf{0}) = \frac{1}{(n+1)q^n} \cdot \left(\frac{q^{\ell(\mathbf{0},\mathbf{y})+1} - 1}{q-1} \right),$$

and in particular, $W_1(\mathbf{0}|\mathbf{0}) = \frac{1}{(n+1)q^n} \cdot \left(\frac{q^{n+1}-1}{q-1}\right)$. Now, given N traces $\mathbf{y}_1, \ldots, \mathbf{y}_N$ generated by passing $\mathbf{0}$ through W_1, let $N_{\mathbf{u}}$ be the random variable denoting the number of occurrences of $\mathbf{u} \in \mathcal{X}^n$ among the traces. It then follows via an application of Hoeffding's inequality [9] (see also [10, Thm. 2.2.6]) that for any $\epsilon > 0$,

$$\Pr\left[N_{\mathbf{0}} \geq N \left(\frac{1}{(n+1)q^n} \cdot \left(\frac{q^{n+1}-1}{q-1} \right) - \epsilon \right) \right] \geq 1 - e^{-2N\epsilon^2}. \tag{1}$$

Furthermore, via a union bound, and using the fact that $W_1(\mathbf{y}|\mathbf{0})$ is increasing in $\ell(\mathbf{0}, \mathbf{y})$, we have for any $\epsilon > 0$ that

$$\Pr\left[N_{\mathbf{y}} \leq N \left(\frac{1}{(n+1)q^n} \cdot \left(\frac{q^n-1}{q-1} \right) + \epsilon \right), \text{ for all } \mathbf{y} \neq \mathbf{0} \right] \geq 1 - (q^n - 1) \cdot e^{-2N\epsilon^2}. \tag{2}$$

Hence, via a union bound, we see that by picking $\epsilon < \frac{1}{2(n+1)}$,

$$\Pr\left[N_{\mathbf{0}} \geq N_{\mathbf{y}}, \text{ for all } \mathbf{y} \neq \mathbf{0} \right] \geq 1 - e^{-\frac{N}{2(n+1)^2} + n \ln q}. \tag{3}$$

Thus, by picking $N \geq 2(n+1)^2 \cdot \left(\ln\left(\frac{1}{\delta}\right) + n \ln q\right)$, we obtain that

$$\Pr\left[\text{FINDMODE}(\mathbf{y}_1, \ldots, \mathbf{y}_N) \neq \mathbf{0} \mid \mathbf{0} \right] \geq 1 - \delta,$$

yielding the statement of the lemma. □

We next obtain an asymptotic lower bound on $T_\delta(n)$.

Theorem 5. *For any $\delta \in (0,1)$, we have $T_\delta(n) = \Omega(n)$.*

Proof. Via Lemma 2, it suffices to characterize $d_n(W_1)$. By symmetry, we have

$$d_n(W_1) = \min_{\mathbf{u} \neq \mathbf{0}} d_{\text{TV}}\left(W_1(\cdot|\mathbf{0}), W_1(\cdot|\mathbf{u}) \right).$$

Now, for any $\mathbf{u} \neq \mathbf{0}$, we have via Lemma 3 that

$$d_{\text{TV}}\left(W_1(\cdot|\mathbf{0}), W_1(\cdot|\mathbf{u})\right) = \frac{q^{-n+1}}{2(n+1)(q-1)} \cdot \sum_{\mathbf{y} \in \mathcal{X}^n} \left| q^{\ell(\mathbf{0},\mathbf{y})} - q^{\ell(\mathbf{u},\mathbf{y})} \right|.$$

Intuitively, each term in the summand above is small if $\ell(\mathbf{0}, \mathbf{y})$ is close to $\ell(\mathbf{u}, \mathbf{y})$. In particular, by picking $\mathbf{u} = \mathbf{0}1$, we obtain after some algebraic manipulations that $d_n(W_1) \leq \frac{1}{n+1}$, thereby yielding the statement of the theorem. □

3. TRIMANDEXTEND

We now consider the channel W_2. Our main contributions are the explicit computations of the channel law and bounds on the quantity $d_n(W_2)$, which naturally lead to bounds on the trace complexity over W_2.

For $\mathbf{u}, \mathbf{y} \in \mathcal{X}^n$, let $L = L(\mathbf{u}, \mathbf{y})$ denote the location of the first disagreement, counting from the left, between $\mathbf{u}$ and $\mathbf{y}$, and let $R = R(\mathbf{u}, \mathbf{y})$ denote the location of the first disagreement, counting from the right (the arguments in $L(\mathbf{u}, \mathbf{y})$ and $R(\mathbf{u}, \mathbf{y})$ are dropped when the sequences are clear from the context).

Lemma 6. *For any $\mathbf{u} \in \mathcal{X}^n$, we have that*

$$W_2(\mathbf{u}|\mathbf{u}) = \frac{2}{(n+1)(n+2)(1-q^{-1})} \cdot \left(\frac{1-q^{-n-1}}{1-q^{-1}} - (n+1)q^{-n-1} \right).$$

Proof. In what follows, let the iterables $r_1, r_2 \geq 0$, with $r_1 + r_2 \leq n$ denote the realizations of the random variables R_1, R_2, respectively, in the definition of the channel W_2. We then have

$$\begin{aligned} W_2(\mathbf{u}|\mathbf{u}) &= \frac{2}{(n+1)(n+2)} \cdot \sum_{r_1=0}^{n} \sum_{r_2=0}^{n-r_1} q^{-r_1-r_2} \\ &= \frac{2}{(n+1)(n+2)} \cdot \sum_{r_1=0}^{n} q^{-r_1} \cdot \left(\frac{1-q^{-(n-r_1+1)}}{1-q^{-1}} \right) \\ &= \frac{2}{(n+1)(n+2)(1-q^{-1})} \cdot \left(\frac{1-q^{-n-1}}{1-q^{-1}} - (n+1)q^{-n-1} \right), \end{aligned}$$

thereby giving rise to the statement of the lemma. □

More generally, the following lemma holds.

Lemma 7. *For any $\mathbf{u}, \mathbf{y} \in \mathcal{X}^n$, we have*

$$\begin{aligned} W_2(\mathbf{y}|\mathbf{u}) = \frac{2}{(n+1)(n+2)(1-q^{-1})} \cdot \Bigg[& q^{-n+R-1-L} \cdot \left(\frac{1-q^{L-R}}{1-q^{-1}} \right) + q^{-R} \cdot \left(\frac{1-q^{-n+R-1}}{1-q^{-1}} \right) \\ & - (n-R+L)q^{-n-1} + q^{-n+L-1} \cdot \left(\frac{1-q^{-L}}{1-q^{-1}} \right) \Bigg]. \end{aligned}$$

Proof. Let the iterables $r_1, r_2 \geq 0$, with $r_1 + r_2 \leq n$ denote the realizations of the random variables R_1, R_2, respectively, in the definition of the channel W_2. We split the evaluation of $W_2(\mathbf{y}|\mathbf{u})$ into

the following three summations:

$$W_2(\mathbf{y}|\mathbf{u}) = \underbrace{\sum_{\substack{r_2 \geq n-R+1,\\ r_1 \geq L}} P_{R_1,R_2}(r_1,r_2) \cdot W_2(\mathbf{y}|\mathbf{u},r_1,r_2)}_{:=\Gamma_1} + \underbrace{\sum_{r_1 \geq R} P_{R_1,R_2}(r_1,r_2) \cdot W_2(\mathbf{y}|\mathbf{u},r_1,r_2)}_{:=\Gamma_2}$$
$$+ \underbrace{\sum_{r_2 \geq n-L+1} P_{R_1,R_2}(r_1,r_2) \cdot W_2(\mathbf{y}|\mathbf{u},r_1,r_2)}_{:=\Gamma_3}.$$

First, consider the term Γ_1. We have

$$\begin{aligned}
\Gamma_1 &= \frac{2}{(n+1)(n+2)} \cdot \sum_{r_1=L}^{n} \sum_{r_2=n-R+1}^{n-r_1} q^{-r_1-r_2} \\
&= \frac{2}{(n+1)(n+2)} \cdot \sum_{r_1=L}^{R-1} q^{-r_1} \cdot q^{-n+R-1} \left(\frac{1-q^{r_1-R}}{1-q^{-1}} \right) \\
&= \frac{2q^{-n+R-1}}{(n+1)(n+2)(1-q^{-1})} \cdot \left[\sum_{r_1=L}^{R-1} q^{-r_1} - \sum_{r_1=L}^{R-1} q^{-R} \right] \\
&= \frac{2q^{-n+R-1}}{(n+1)(n+2)(1-q^{-1})} \cdot \left[q^{-L} \cdot \left(\frac{1-q^{L-R}}{1-q^{-1}} \right) - (R-L)q^{-R} \right].
\end{aligned}$$

Next, we take up the term Γ_2. We have

$$\begin{aligned}
\Gamma_2 &= \frac{2}{(n+1)(n+2)} \cdot \sum_{r_1=R}^{n} \sum_{r_2=0}^{n-r_1} q^{-r_1-r_2} \\
&= \frac{2}{(n+1)(n+2)} \cdot \sum_{r_1=R}^{n} q^{-r_1} \cdot \left(\frac{1-q^{-n+r_1-1}}{1-q^{-1}} \right) \\
&= \frac{2}{(n+1)(n+2)(1-q^{-1})} \cdot \left[q^{-R} \cdot \left(\frac{1-q^{-n+R-1}}{1-q^{-1}} \right) - (n-R+1)q^{-n-1} \right].
\end{aligned}$$

Finally, consider the term Γ_3. We have

$$\begin{aligned}
\Gamma_3 &= \frac{2}{(n+1)(n+2)} \cdot \sum_{r_2=n-L+1}^{n} \sum_{r_1=0}^{n-r_2} q^{-r_1-r_2} \\
&= \frac{2}{(n+1)(n+2)} \cdot \sum_{r_2=n-L+1}^{n} q^{-r_2} \cdot \left(\frac{1-q^{-n+r_2-1}}{1-q^{-1}} \right) \\
&= \frac{2}{(n+1)(n+2)(1-q^{-1})} \cdot \left[q^{-n+L-1} \cdot \left(\frac{1-q^{-L}}{1-q^{-1}} \right) - Lq^{-n-1} \right].
\end{aligned}$$

Putting everything together yields the statement of the lemma. □

Observe that Lemma 7 gives rise to Lemma 6 as a corollary, by plugging in $L = R = 0$. The characterization above again directly gives rise to a simple sequence reconstruction algorithm, via FINDMODE.

Theorem 8. *For any $\delta \in (0,1)$, we have $T_\delta(n) = O(n^4)$.*

Proof. Recall from Lemma 1 that it suffices to consider the trace complexity $T_\delta(n;\mathbf{0})$. Now, via Lemma 7, we have that for any $\mathbf{y} \neq \mathbf{0}$,

$$W_2(\mathbf{y}|\mathbf{0}) = \frac{2}{(n+1)(n+2)(1-q^{-1})} \cdot \Bigg[q^{-n+R-1-L} \cdot \left(\frac{1-q^{L-R}}{1-q^{-1}}\right) + q^{-R} \cdot \left(\frac{1-q^{-n+R-1}}{1-q^{-1}}\right) - (n-R+L)q^{-n-1} + q^{-n+L-1} \cdot \left(\frac{1-q^{-L}}{1-q^{-1}}\right) \Bigg],$$

where L and R are, respectively, the locations of the first 1, counting from the left, and the first 1, counting from the right, in $\mathbf{y}$. In particular, $W_2(\mathbf{0}|\mathbf{0}) = \frac{2}{(n+1)(n+2)(1-q^{-1})} \cdot \left(\frac{1-q^{-n-1}}{1-q^{-1}} - (n+1)q^{-n-1}\right)$, via Lemma 6. Observe that since $1 \leq L \leq n$ and $1 \leq R \leq n$, we have that

$$\begin{aligned} W_2(\mathbf{0}|\mathbf{0}) - W_2(\mathbf{y}|\mathbf{0}) &\geq \frac{2}{(n+1)(n+2)(1-q^{-1})^2} \cdot \left(2q^{-1} + q^{-2} + (1-4q)q^{-n-2}\right) \\ &\geq \frac{2}{(n+1)(n+2)(1-q^{-1})^2} \cdot \left(2q^{-1} + q^{-2} + (1-4q)q^{-3}\right) =: c_q(n), \end{aligned} \tag{4}$$

for all $n \geq 1$. It can easily be checked that the right-hand side of (4) is strictly positive, for $q \geq 2$.

Now, given N traces $\mathbf{y}_1, \ldots, \mathbf{y}_N$ generated by passing $\mathbf{0}$ through W_1, let $N_{\mathbf{u}}$ be the random variable denoting the number of occurrences of $\mathbf{u} \in \mathcal{X}^n$ among the traces. Via arguments entirely similar to that in Theorem 4, we have for any $\epsilon > 0$ that

$$\Pr\left[N_{\mathbf{0}} \geq N\left(W_2(\mathbf{0}|\mathbf{0}) - \epsilon\right)\right] \geq 1 - e^{-2N\epsilon^2}. \tag{5}$$

Furthermore, via a union bound, we have for any $\epsilon > 0$ that

$$\Pr\left[N_{\mathbf{y}} \leq N\left(W_2(\mathbf{y}|\mathbf{0}) + \epsilon\right), \text{ for all } \mathbf{y} \neq \mathbf{0}\right] \geq 1 - (q^n - 1) \cdot e^{-2N\epsilon^2}. \tag{6}$$

Hence, via a union bound, we see from (4) that by picking $\epsilon < c_q(n)/2$,

$$\Pr\left[N_{\mathbf{0}} \geq N_{\mathbf{y}}, \text{ for all } \mathbf{y} \neq \mathbf{0}\right] \geq 1 - e^{-2N(c_q(n))^2 + n\ln q}. \tag{7}$$

Thus, by picking $N \geq \frac{\ln\left(\frac{1}{\delta}\right) + n\ln q}{2(c_q(n))^2}$, we obtain that

$$\Pr\left[\textsc{FindMode}(\mathbf{y}_1, \ldots, \mathbf{y}_N) \neq \mathbf{0} \mid \mathbf{0}\right] \geq 1 - \delta,$$

which yields the statement of the lemma, since $\frac{1}{(c_q(n))^2} = O(n^4)$. □

In a manner entirely analogous to Theorem 5, we shall obtain an asymptotic lower bound on $T_\delta(n)$, over W_2.

Theorem 9. *For any $\delta \in (0,1)$, we have that $T_\delta(n) = \Omega(n^2)$, if n is odd.*

Proof. We proceed similar to the proof of Theorem 5. By symmetry once again, we have

$$d_n(W_2) = \min_{\mathbf{u} \neq \mathbf{0}} d_{\text{TV}}\left(W_2(\cdot|\mathbf{0}), W_2(\cdot|\mathbf{u})\right).$$

For ease of reading, we set $L_{\mathbf{u}} := L(\mathbf{u}, \mathbf{y})$, when the sequence $\mathbf{y}$ is clear from the context. Now, for any fixed $\mathbf{u} \in \mathcal{X}^n$, we have from Lemma 7 that

$$\begin{aligned} (n+1)(n+2)(1-q^{-1}) \cdot d_{\text{TV}}\left(W_2(\cdot|\mathbf{0}), W_2(\cdot|\mathbf{u})\right) &= \sum_{\mathbf{y} \in \mathcal{X}^n} \left|\beta_1(\mathbf{y}) + \beta_2(\mathbf{y}) + \beta_3(\mathbf{y}) + \beta_4(\mathbf{y})\right| \\ &\leq \sum_{\mathbf{y} \in \mathcal{X}^n} \sum_{i=1}^{4} \left|\beta_i(\mathbf{y})\right|, \end{aligned} \tag{8}$$

where

$$\begin{aligned}
\beta_1(\mathbf{y}) &:= q^{-n-1}\cdot\left(q^{R_{\mathbf{0}}-L_{\mathbf{0}}}\cdot\left(\frac{1-q^{L_{\mathbf{0}}-R_{\mathbf{0}}}}{1-q^{-1}}\right)-q^{R_{\mathbf{u}}-L_{\mathbf{u}}}\cdot\left(\frac{1-q^{L_{\mathbf{u}}-R_{\mathbf{u}}}}{1-q^{-1}}\right)\right)\\
&= \frac{q^{-n-1}}{1-q^{-1}}\cdot\left(q^{R_{\mathbf{0}}-L_{\mathbf{0}}}-q^{R_{\mathbf{u}}-L_{\mathbf{u}}}\right);\\
\beta_2(\mathbf{y}) &:= q^{-R_{\mathbf{0}}}\cdot\left(\frac{1-q^{-n+R_{\mathbf{0}}-1}}{1-q^{-1}}\right)-q^{-R_{\mathbf{u}}}\cdot\left(\frac{1-q^{-n+R_{\mathbf{u}}-1}}{1-q^{-1}}\right);\\
\beta_3(\mathbf{y}) &:= q^{-n-1}\cdot\left(L_{\mathbf{u}}-R_{\mathbf{u}}+R_{\mathbf{0}}-L_{\mathbf{0}}\right);\text{ and}\\
\beta_4(\mathbf{y}) &:= q^{-n-1}\cdot\left(q^{L_{\mathbf{0}}}\cdot\left(\frac{1-q^{-L_{\mathbf{0}}}}{1-q^{-1}}\right)-q^{L_{\mathbf{u}}}\cdot\left(\frac{1-q^{-L_{\mathbf{u}}}}{1-q^{-1}}\right)\right).
\end{aligned}$$

Intuitively, the expression on the right in (8) is small, if $L_{\mathbf{0}}$ and $L_{\mathbf{u}}$ (resp. $R_{\mathbf{0}}$ and $R_{\mathbf{u}}$) are close for many sequences $\mathbf{y}$. Following this intuition, we pick $\mathbf{u} = 0^{(n-1)/2}10^{(n-1)/2}$. Let $h(\mathbf{y}) := \sum_{i=1}^{4}|\beta_i(\mathbf{y})|$.

Via the definition of $\mathbf{u}$, we see that if $\mathbf{y}$ is such that $L_{\mathbf{0}} < \frac{n+1}{2}$ and $R_{\mathbf{0}} < \frac{n+1}{2}$, then $h(\mathbf{y}) = 0$. Furthermore, we claim that

$$\sum_{\mathbf{y}:L_{\mathbf{0}}\leq\frac{n+1}{2},\ R_{\mathbf{0}}\geq\frac{n+1}{2}} h(\mathbf{y}) = \sum_{\mathbf{y}:L_{\mathbf{0}}\geq\frac{n+1}{2},\ R_{\mathbf{0}}\leq\frac{n+1}{2}} h(\mathbf{y}), \tag{9}$$

and likewise,

$$\sum_{\mathbf{y}:L_{\mathbf{0}}<\frac{n+1}{2},\ R_{\mathbf{0}}<\frac{n+1}{2}} h(\mathbf{y}) = \sum_{\mathbf{y}:L_{\mathbf{0}}>\frac{n+1}{2},\ R_{\mathbf{0}}>\frac{n+1}{2}} h(\mathbf{y}) = 0. \tag{10}$$

To see why (9) and (10) hold, we observe that every summand on the left-hand side of (9) (resp. (10)) can be mapped uniquely to a summand on the right-hand side of (9) (resp. (10)), via the mapping $(L_{\mathbf{0}}, R_{\mathbf{0}}) \mapsto (n+1-R_{\mathbf{0}}, n+1-L_{\mathbf{0}})$, and vice-versa. Hence, it follows that

$$\Delta := (n+1)(n+2)(1-q^{-1})\cdot d_{\text{TV}}\left(W_2(\cdot|\mathbf{0}), W_2(\cdot|\mathbf{u})\right) = 2\cdot\sum_{\mathbf{y}:L_{\mathbf{0}}\leq\frac{n+1}{2},\ R_{\mathbf{0}}\geq\frac{n+1}{2}} h(\mathbf{y}). \tag{11}$$

We hence restrict attention to the expression $\sum_{\mathbf{y}:L_{\mathbf{0}}\leq\frac{n+1}{2},\ R_{\mathbf{0}}\geq\frac{n+1}{2}} h(\mathbf{y})$ above. In this setting, we have $R_{\mathbf{u}} \geq \frac{n+1}{2}$. Since for any fixed $(\ell_{\mathbf{0}}, r_{\mathbf{0}})$ pair, there are $(q-1)^2q^{n-\ell_{\mathbf{0}}-r_{\mathbf{0}}} < q^{n-\ell_{\mathbf{0}}-r_{\mathbf{0}}+2}$ sequences $\mathbf{y}$ with $L_{\mathbf{0}} = \ell_{\mathbf{0}}$ and $R_{\mathbf{0}} = r_{\mathbf{0}}$, it follows from (11) that

$$\Delta = 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n+1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{n+1-L_{\mathbf{0}}} q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot h(L_{\mathbf{0}}, R_{\mathbf{0}}, L_{\mathbf{u}}, R_{\mathbf{u}}), \tag{12}$$

where, with some abuse of notation, we have set $h(\mathbf{y}) = h(L_{\mathbf{0}}, R_{\mathbf{0}}, L_{\mathbf{u}}, R_{\mathbf{u}})$, if $\mathbf{y}$ corresponds to the pair $(L_{\mathbf{0}}, R_{\mathbf{0}})$, under the input sequence $\mathbf{0}$, and $(L_{\mathbf{u}}, R_{\mathbf{u}})$, under the input sequence $\mathbf{u}$, respectively. We further write β_1 through β_4 as functions of the tuple $(L_{\mathbf{0}}, R_{\mathbf{0}}, L_{\mathbf{u}}, R_{\mathbf{u}})$, similarly.

Consider first the summation

$$\begin{aligned}\Delta_4 &:= 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n+1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{n+1-L_{\mathbf{0}}} q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot|\beta_4(\mathbf{y})|\\ &= 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{n+1-L_{\mathbf{0}}} q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot|\beta_4(\mathbf{y})| + 2q^{n+2}\cdot\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{n+1-L_{\mathbf{0}}}\left(q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot|\beta_4(\mathbf{y})|\right)\Big|_{L_{\mathbf{0}}=\frac{n+1}{2}}\\ &= 2q^{\cdot}|\beta_4(\mathbf{y})|\Big|_{L_{\mathbf{0}}=\frac{n+1}{2}} = O(1),\end{aligned} \tag{13}$$

where the penultimate equality holds since we have that $L_{\mathbf{0}} = L_{\mathbf{u}}$, when $L_{\mathbf{0}} \leq \frac{n-1}{2}$. Moreover, we have

$$\Delta_3 := 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n+1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{n+1-L_{\mathbf{0}}} q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot|\beta_3(L_{\mathbf{0}},R_{\mathbf{0}},L_{\mathbf{u}},R_{\mathbf{u}})| = O(1). \tag{14}$$

Now, note that

$$\begin{aligned}\Delta_1 &:= 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n+1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{n+1-L_{\mathbf{0}}} q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot|\beta_1|\\ &\leq 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+3}{2}}^{n+1-L_{\mathbf{0}}} q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot|\beta_1| + 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{\frac{n+3}{2}} q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot|\beta_1| + O(1)\\ &\leq 2q\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+3}{2}}^{n+1-L_{\mathbf{0}}} q^{-2L_{\mathbf{0}}-R_{\mathbf{0}}+\frac{n+1}{2}} + 2q\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{\frac{n+3}{2}} q^{-2L_{\mathbf{0}}-R_{\mathbf{0}}+\frac{n+3}{2}} + O(1) = O(1).\end{aligned} \tag{15}$$

Furthermore, we have

$$\begin{aligned}&\Delta_2\\ &:= 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n+1}{2}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{n+1-L_{\mathbf{0}}} q^{-L_{\mathbf{0}}-R_{\mathbf{0}}}\cdot|\beta_2|\\ &\leq 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}} q^{-L_{\mathbf{0}}}\sum_{R_{\mathbf{0}}=\frac{n+3}{2}}^{n+1-L_{\mathbf{0}}} q^{-R_{\mathbf{0}}}\cdot\left(q^{-R_{\mathbf{0}}}\cdot\left(\frac{1-q^{-n+R_{\mathbf{0}}-1}}{1-q^{-1}}\right)+q^{-\frac{n+1}{2}}\cdot\left(\frac{1-q^{-\frac{n+1}{2}}}{1-q^{-1}}\right)\right)\\ &+ 2q^{n+2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}} q^{-L_{\mathbf{0}}}\sum_{R_{\mathbf{0}}=\frac{n+1}{2}}^{\frac{n+3}{2}} q^{-R_{\mathbf{0}}}\cdot\left(q^{-R_{\mathbf{0}}}\cdot\left(\frac{1-q^{-n+R_{\mathbf{0}}-1}}{1-q^{-1}}\right)+q^{-\frac{n+3}{2}}\cdot\left(\frac{1-q^{-\frac{n+3}{2}}}{1-q^{-1}}\right)\right)+O(1)\\ &\leq \frac{2q}{(1-q^{-1})^2}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}} q^{-L_{\mathbf{0}}} + \frac{4q}{(1-q^{-1})}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}} q^{-L_{\mathbf{0}}} + \frac{4q}{(1-q^{-1})}\cdot\sum_{L_{\mathbf{0}}=0}^{\frac{n-1}{2}} q^{-L_{\mathbf{0}}} + O(1) = O(1).\end{aligned} \tag{16}$$

Algorithm 2 Sequence reconstruction algorithm

1: **procedure** TRIMANDFINDMODE($\mathbf{y}_1, \ldots, \mathbf{y}_N$)
2: Let $\overline{\mathbf{y}}_1, \ldots, \overline{\mathbf{y}}_R$ be the collection of traces of length at least n.
3: Trim the sequences $\overline{\mathbf{y}}_i$ to retain their length-n prefixes $\widehat{\mathbf{y}}_i$, $i \in [R]$.
4: Return the sequence occurring most often in $(\widehat{\mathbf{y}}_1, \ldots, \widehat{\mathbf{y}}_R)$.

The inequalities that led to (15) and (16) were obtained via repeated applications of the triangle inequality and simple approximations. Putting together (11)–(16), we see that

$$d_{\mathrm{TV}}\left(W_2(\cdot|\mathbf{0}), W_2(\cdot|\mathbf{u})\right) \leq \frac{1}{(n+1)(n+2)(1-q^{-1})} \sum_{i=1}^{4} \Delta_i = O\left(\frac{1}{(n+1)(n+2)}\right),$$

thereby yielding the statement of the theorem. □

4. SUFFIXEXTEND$_t$(TRIMSUFFIX)

We now turn to the channel W_3. As before, we first present a simple sequence reconstruction algorithm for this channel that succeeds with high probability, thus giving rise to an upper bound on the trace complexity. We first recapitulate the channel law for this channel given in [1, Sec. III-B]. Now, let $\ell(\mathbf{u}, \mathbf{y})$ denote the longest common prefix between strings $\mathbf{u} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{X}^\star$. Note that this notation is an extension of that employed in Section 2. The following lemma then holds:

Lemma 10 ([1], Eq. (3))**.** *For any $\mathbf{u} \in \mathcal{X}^n, \mathbf{y} \in \mathcal{X}^m$, for some $0 \leq m \leq n+t$, we have*

$$W_3(\mathbf{y}|\mathbf{u}) = \frac{1}{(n+1)(t+1)} \cdot \sum_{k=(m-t)_+}^{\ell(\mathbf{u},\mathbf{y})} \frac{1}{q^{m-k}}.$$

We now present a simple sequence reconstruction algorithm, TRIMANDFINDMODE, over W_3, Algorithm 2. The following lemma will be useful to us; let $\ell = \ell(\mathbf{y}) := \ell(\mathbf{0}, \mathbf{y})$, when the sequence $\mathbf{y}$ is clear from the context. Further, let $\lambda := t + \ell - n + 1$.

Lemma 11. *We have that for any $\mathbf{y} \in \mathcal{X}^n$,*

$$\Pr\left[\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{y}}_1|\mathbf{0}\right] = \begin{cases} \frac{q^{-\ell}}{(n+1)(t+1)(q-1)^2} \cdot \left[q^{\lambda+1} - (\lambda+1)q + \lambda\right], \text{ if } \ell \geq n-t, \\ 0, \text{ otherwise.} \end{cases}$$

Proof. Suppose that $\Pr\left[\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{y}}_1|\mathbf{0}\right] > 0$, when $n - \ell(\widehat{\mathbf{y}}_1) \geq t+1$. Then, the length of $\overline{\mathbf{y}}_1$ that was trimmed to obtain $\widehat{\mathbf{y}}_1$ is at most $n-1$, leading to a contradiction. In the case when $\ell(\widehat{\mathbf{y}}_1) \geq n-t$, we have

$$\begin{aligned}
\Pr\left[\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{y}}_1|\mathbf{0}\right] &= \sum_{r=n-\ell}^{t} P_R(r) \cdot \Pr[E \geq r] \cdot q^{r-n} \\
&= \frac{1}{(n+1)(t+1)} \cdot \sum_{r=n-\ell}^{t} (t-r+1) \cdot q^{r-n} \\
&= \frac{q^{-n+t+1}}{(n+1)(t+1)} \cdot \sum_{r=1}^{t+\ell-n+1} r \cdot q^{-r} \\
&= \frac{q^{-\ell}}{(n+1)(t+1)} \cdot \left(\frac{q^{t+\ell-n+2} - (t+\ell-n+2)q + (t+\ell-n+1)}{(q-1)^2}\right) \\
&= \frac{q^{-\ell}}{(n+1)(t+1)(q-1)^2} \cdot \left[q^{\lambda+1} - (\lambda+1)q + \lambda\right],
\end{aligned}$$

thereby giving us the statement of the lemma. □

We next present a lemma that shows argues that $\Pr\left[\widehat{\mathbf{Y}}_1 = \mathbf{0}|\mathbf{0}\right]$ is strictly larger than $\Pr\left[\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{y}}_1|\mathbf{0}\right]$, for all $\widehat{\mathbf{y}}_1 \neq \mathbf{0}$. To this end, we note from Lemma 11 that it suffices to show that $\Pr\left[\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{y}}_1|\mathbf{0}\right]$ is increasing in ℓ.

Lemma 12. *For any $\widehat{\mathbf{y}}_1 \in \mathcal{X}^n$ such that $\widehat{\mathbf{y}}_1 \neq \mathbf{0}$, we have $\Pr\left[\widehat{\mathbf{Y}}_1 = \mathbf{0}|\mathbf{0}\right] > \Pr\left[\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{y}}_1|\mathbf{0}\right]$. Furthermore,*

$$\Pr\left[\widehat{\mathbf{Y}}_1 = \mathbf{0}|\mathbf{0}\right] - \max_{\widehat{\mathbf{y}}_1 \neq \mathbf{0}} \Pr\left[\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{y}}_1|\mathbf{0}\right] = \frac{q^{-n}}{(n+1)}.$$

Proof. For the first statement of the lemma, following Lemma 11, it suffices to show that

$$g(\ell) := q^{-\ell} \cdot (t + \ell - n + 1 - (t + \ell - n + 2)q) - q^{-\ell+1} \cdot (t + \ell - n - (t + \ell - n + 1)q)$$

satisfies $g(\ell) > 0$, for all $n - t \leq \ell \leq n$. Indeed, observe that

$$g(\ell) = q^{-\ell} \cdot (t + \ell - n + 1) \cdot (q-1)^2 > 0,$$

for all $q > 1$, $n - t \leq \ell \leq n$. The second statement of the lemma holds by observing that $\max_{\widehat{\mathbf{y}}_1 \neq \mathbf{0}} \Pr\left[\widehat{\mathbf{Y}}_1 = \widehat{\mathbf{y}}_1|\mathbf{0}\right]$ is attained by any $\widehat{\mathbf{y}}_1$ with $\ell(\widehat{\mathbf{y}}_1) = n - 1$. □

Let $\overline{c}_q(n) := \frac{q^{-n}}{(n+1)}$. Via arguments entirely analogous to the proofs of Theorems 4 and 8, we obtain the following theorem:

Theorem 13. *For any $\delta \in (0,1)$, we have that $\Pr\left[\textsc{TrimAndFindMode}(\mathbf{y}_1, \ldots, \mathbf{y}_N) \neq \mathbf{0} \mid \mathbf{0}\right] \geq 1 - \delta$, when $N \geq \frac{\ln\left(\frac{1}{\delta}\right) + n \ln q}{2(\overline{c}_q(n))^2}$. In particular, we have that $T_\delta(n) = O(n^3)$.*

Next, we state an asymptotic lower bound on $T_\delta(n)$, in a manner similar to Theorems 5 and 9.

Theorem 14. *For any $\delta \in (0,1)$, we have that $T_\delta(n) = \Omega(n)$.*

Proof. Via Lemma 2, it suffices to characterize $d_n(W_3)$. Once again, by symmetry

$$d_n(W_3) = \min_{\mathbf{u} \neq \mathbf{0}} d_{\text{TV}}\left(W_3(\cdot|\mathbf{0}), W_3(\cdot|\mathbf{u})\right).$$

Now, for any $\mathbf{u} \neq \mathbf{0}$, we have via Lemma 10 that

$$d_{\text{TV}}\left(W_3(\cdot|\mathbf{0}), W_3(\cdot|\mathbf{u})\right) = \frac{1}{2(n+1)(t+1)} \cdot \sum_{\mathbf{y} \in \mathcal{X}^n} \left| \sum_{k=(|\mathbf{y}|-t)_+}^{\ell_{\mathbf{0}}} q^{k-|\mathbf{y}|} - \sum_{k=(|\mathbf{y}|-t)_+}^{\ell_{\mathbf{u}}} q^{k-|\mathbf{y}|} \right|,$$

where $\ell_{\mathbf{0}}$ and $\ell_{\mathbf{u}}$, respectively are shorthand for $\ell(\mathbf{0}, \mathbf{y})$ and $\ell(\mathbf{u}, \mathbf{y})$, when $\mathbf{y}$ is clear from the context. Intuitively, each term in the summand above is small if $\ell(\mathbf{0}, \mathbf{y})$ is close to $\ell(\mathbf{u}, \mathbf{y})$. In particular, by picking $\mathbf{u} = \mathbf{0}1$, we obtain after some algebraic manipulations that $d_n(W_3) \leq \frac{1}{2(n+1)(t+1)}$, thereby yielding the statement of the theorem. □

REFERENCES

[1] V. Bhardwaj, P. A. Pevzner, C. Rashtchian, and Y. Safonova, "Trace reconstruction problems in computational biology," *IEEE Transactions on Information Theory*, vol. 67, no. 6, pp. 3295–3314, 2021.

[2] V. Levenshtein, "Efficient reconstruction of sequences," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 2–22, 2001.

[3] V. I. Levenshtein, "Efficient reconstruction of sequences from their subsequences or supersequences," *Journal of Combinatorial Theory, Series A*, vol. 93, no. 2, pp. 310–332, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0097316500930814

[4] T. Batu, S. Kannan, S. Khanna, and A. McGregor, "Reconstructing strings from random traces," in *Proceedings of the Fifteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, ser. SODA '04. USA: Society for Industrial and Applied Mathematics, 2004, p. 910–918.

[5] F. Nazarov and Y. Peres, "Trace reconstruction with $\exp(o(n^{1/3}))$ samples," in *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2017. New York, NY, USA: Association for Computing Machinery, 2017, p. 1042–1046. [Online]. Available: https://doi.org/10.1145/3055399.3055494

[6] Y. Peres and A. Zhai, "Average-case reconstruction for the deletion channel: Subpolynomially many traces suffice," in *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, 2017, pp. 228–239.

[7] Z. Chase, "Separating words and trace reconstruction," in *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, ser. STOC 2021. New York, NY, USA: Association for Computing Machinery, 2021, p. 21–31. [Online]. Available: https://doi.org/10.1145/3406325.3451118

[8] ——, "New lower bounds for trace reconstruction," *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 57, no. 2, pp. 627 – 643, 2021. [Online]. Available: https://doi.org/10.1214/20-AIHP1089

[9] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol. 58, no. 301, pp. 13–30, 1963. [Online]. Available: http://www.jstor.org/stable/2282952

[10] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*, ser. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.